Label-Sensitive Deep Metric Learning for Facial Age Estimation

Hao Liu, Jiwen Lu, *Senior Member, IEEE*, Jianjiang Feng, *Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

***Abstract*— In this paper, we present a label-sensitive deep metric learning (LSDML) approach for facial age estimation. Motivated by the fact that human age labels are chronologically correlated, our proposed LSDML aims to seek a series of hierarchical nonlinear transformations by deep residual network to project face samples to a latent common space, where the similarity of face pairs is equivalently isotonic to the age difference in a ranking-preserving manner. Since traversal access to total negative samples catastrophically costs and leads to suboptimal, our model learns to mine hard meaningful samples in parallel to learning feature similarity, so that the local manifold of face samples is preserved in the transformed subspace. To better improve the performance on the data set that contains few labeled samples, we further extend our LSDML to a multi-source LSDML method, which aims at maximizing the cross-population correlation of different face aging data sets. Extensive experimental results on four benchmarking data sets show the effectiveness of our proposed approach.**

***Index Terms*— Facial age estimation, metric learning, deep learning, residual network, biometrics.**

I. Introduction

FACIAL age estimation has been widely used in various applications such as facial attributes detection, visual advertisements and biometrics [1]–[5], which attempts to predict exact age values for given facial images. While extensive efforts have been devoted to facial age estimation areas, the performance is still not satisfied in practice due to the variances of diverse facial expressions, aspect ratios, clutter background and partial occlusions, especially when face images were captured in wild conditions (as shown in Fig. 1). Moreover, facial age estimation usually encounters predicting ambiguity due to correlated age classes. In this paper, we consider exploiting the label correlation in a deeply embedded subspace, where the feature similarity of face pairs is smoothly sensitive to the age difference values.

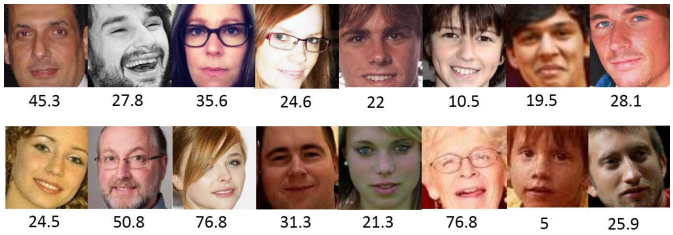


Fig. 1. The cropped face samples from the uncontrolled face aging dataset, where the number below each face image is the apparent age value. We see that the face samples in this dataset usually undergo challenging situations due to diverse facial expressions, aspect ratios, cluttered background and partial occlusions.

Recently, many approaches have been proposed to improve the performance for facial age estimation [6]–[10], which can be roughly divided into two categories: face representation-based [6], [7], [11] and age estimator-based [9], [12], [13]. However, feature descriptors for face representation employed in previous methods are usually hand-crafted, which may loss crucial information and requires expert knowledge by hand. To address this, feature learning [14], [15] has been proposed, which learns discriminative filters directly from raw pixels for robust age-related face representation. While these data-driven methods outperform hand-crafted features, they are still not satisfied in practical applications because these simple linear filters are not powerful enough to exploit the nonlinear relationship between face samples and age labels, especially when face samples were captured in the uncontrolled environments. To address this limitation, deep learning-based techniques [16]–[19] have been applied to learn a series of hierarchical nonlinear mappings via stacked deep neural networks to transform the input face images to the age label space. For example, Yi *et al.* [16] developed a multi-scale framework to learn deep face representations to predict the age value with the auxiliary gender and ethnicity information. Niu *et al.* [19] proposed an ordinal regression method via deep convolutional neural networks to exploit the age order information. While very competitive performance has been obtained, learning parameters for the leveraged deep architectures catastrophically costs and might lead to suboptimal due to the imbalanced positive and negative samples during training process.

Manuscript received April 24, 2017; revised July 6, 2017 and August 11, 2017; accepted August 21, 2017. Date of publication August 29, 2017; date of current version November 28, 2017. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001001, in part by the National Natural Science Foundation of China under Grant 61672306, Grant 61572271, Grant 61527808, Grant 61373074, and Grant 61373090, in part by the National 1000 Young Talents Plan Program, in part by the National Basic Research Program of China under Grant 2014CB349304, in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564, in part by the Ministry of Education of China under Grant 20120002110033, and in part by the Tsinghua University Initiative Scientific Research Program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Karthik Nandakumar. *(Corresponding author: Jiwen Lu.)*

The authors are with the Department of Automation, Tsinghua University, State Key Lab of Intelligent Technologies and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing 100084, China (e-mail: h-liu14@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn; jfeng@tsinghua.edu.cn; jzhou@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2017.2746062

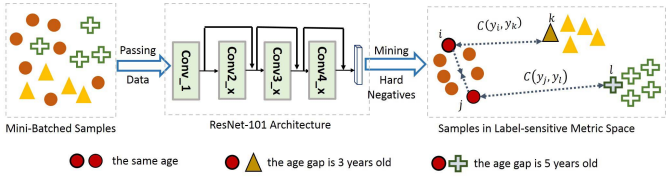


Fig. 2. The work-flow of our proposed LSDML. Specifically, we first pass a mini-batch forward the designed deep residual network, and then select a subset of hard quadruplets. Having obtained the mined hard samples, our model enforces the relationship within each quadruplet in a transformed subspace, where the hard positive samples are pushed as close as possible, and at the same time hard negative samples are pushed larger than a label-sensitive threshold. Note that both the tasks of mining hard examples and learning feature similarity are jointly optimized to reinforce our model.

It is widely observed that effective data sampling techniques [20], [21] are crucial to ensure the quality and efficiency for robust feature similarity. Inspired by this, we propose a label-sensitive deep metric learning (LSDML) method for facial age estimation, which jointly optimizes both procedures of mining hard examples and learning discriminative metric in a unified deep architecture. Fig. 2 illustrates the work-flow of the proposed LSDML. Specifically, our LSDML aims to learn a series hierarchical nonlinear mappings via deep residual network to transform raw face images to a latent common subspace, where the inter-class compactness and intra-class separability are exploited. Moreover, our model learns to exploit the label correlation in the transformed subspace, so that the distance of face pairs with different age differences is smoothly measured according to the degree of age difference. To make the process of feature similarity learning more efficient, we propose a sampling strategy to select meaningful hard examples on parameter update during the back-propagation process. As a result, both tasks of feature similarity learning and hard samples mining are jointly optimized in an end-to-end manner. To further enhance the discriminativeness of the learned metric, we present a multi-source LSDML method by maximizing cross-population correlation of multi-source datasets, which considerably addresses the limitation of few labeled training samples in some face aging datasets. Experimental results on four face benchmarking datasets in both controlled and uncontrolled environments show the superior performance in comparisons to the state-of-the-art facial age estimation methods.

The contributions of this work are summarized as follows:

1) We propose a label-sensitive deep metric learning (LSDML) method to learn a discriminative distance metric under the paradigm of deep residual network, where the inter-class separability, intra-class compactness, and label correlation of age classes are simultaneously exploited to faithfully characterize the true feature similarity for face aging data.
2) To make the learned similarity more effective and efficient, we develop a hard example mining strategy, so that the local manifold structure of input data points is exploited in the learned metric space. Moreover, we jointly optimize both the tasks of learning label-sensitive feature similarity and mining hard samples directly from image pixels, which exploits the complementary information for both tasks to reinforce our model in an end-to-end manner.
3) We extend our proposed LSDML to a multi-source LSDML method, which attempts to maximize the cross-population correlation between different face aging datasets to enhance the robustness and discriminativeness of the learned feature similarity.

II. Related Work

In this section, we review two related works including facial age estimation and deep metric learning.

A. Facial Age Estimation

Existing facial age estimation methods [9], [12], [22]–[24] can be categorized into two classes: face representation-based and age estimator-based. The representative face representation-based methods include the holistic subspace feature [25], [26], local binary pattern (LBP) [7] and the bio-inspired feature (BIF) [11]. The representative age estimator-based methods include age pattern regression [22], multi-task warped Gaussian process (MTWGP) [23], ordinal hyperplane ranking (OHRANK) [9] and label distribution learning (LDL) [12]. However, the face descriptors employed in previous methods were designed by hand, which requires strong prior knowledge and even performs degraded performance when face samples were captured in wild conditions. To overcome this limitation, feature learning-based methods [15], [24], [27] have been proposed to learn robust face descriptors for age estimation. For example, Fu *et al.* [14] proposed a holistic feature learning method by using a discriminative manifold learning technique. Lu *et al.* [15] developed a cost-sensitive local binary feature learning method for facial age estimation. However, the performance is still far from the satisfactory, because the employed linear feature filters are not powerful enough to exploit the nonlinear relationship of face aging data, especially when faces were exposed to large variances of facial expressions and even partial occlusions (*e.g.*, make-up, wearing glasses, etc.). To address this nonlinear issue, deep learning has been applied for facial age estimation [19], [28]–[32], which aims to learn a set of nonlinear feature transformations and achieves the nonlinear relationship between face samples and age labels. For example, Levi and Hassner [29] proposed a multi-task framework via deep convolutional neural network to jointly address the age and gender classification in a unified deep learning framework. Yang *et al.* [33] employed deep scattering transform networks (DeepRank) to predict ages via category-wise rankers. Niu *et al.* [19] developed an ordinal regression by the convolutional neural network (OR-CNN) method with multiple binary outputs for age estimation. While promising performance has been obtained, these methods ignore to take advantages of sampling effective data points, so that the training procedure is time-consuming and may incur suboptimal because of the unbalances positive and negative samples. To circumvent this problem, we propose a label-sensitive deep metric learning method to automatically learn to select

hard samples on parameter update during back-propagation procedure. With the learned metric, the label correlation of aging pattern is exploited in the embedded feature space, and at the same time hard meaningful examples are captured, which makes the learned similarity metric more robust and discriminative£ and leads to fast convergence.

B. Deep Metric Learning

Recent years have witnessed that deep metric learning has received much attention in the research fields of machine learning and computer vision due to its superior performance [34]–[40], which aims to optimize both tasks of learning nonlinear subspace and embedded feature similarity by leveraging different deep architectures. For example, Bromley *et at.* [35] and Chopra *et al.* [36] trained Siamese networks via deep neural networks for signature and face verification. Hu *et al.* [37], [38] proposed two deep metric learning methods including the discriminative distance metric learning (DDML) and the deep transfer metric learning (DTML), which aim to transform similar input objects on a manifold and dissimilar objects apart from each other by the triplet [41] loss. To further mine the high-order and structural relations of the local manifold structure for input data points, Song *et al.* [39] and Huang *et al.* [40] developed deep feature embedding approaches by leveraging the quadruplet-based comparisons [42], which significantly promotes the accuracy for image classification and retrieval. However, these deep metric learning methods ignore to explicitly exploit the label correlation for the specific classes, which cannot be directly applied on facial age estimation because the age labels of target variables exhibit a natural ordering in the real-world applications.

In contrast to previous methods, we propose a label-sensitive deep metric learning method to measure the face samples in a transformed subspace depending on the smoothing degree of age difference, in addition to optimizing both the intra-class compactness and inter-class separability under a unified deep neural network framework. To overcome the limitation of full access to negative samples, we develop a hard example mining strategy to automatically discover the semantic and meaningful violated samples during the training procedure, which efficiently reduces the computational cost and leads to fast convergence. Besides, we introduce a multi-source LSDML method to maximize the cross-population correlation between different datasets, which circumvents the problems of missing labels and unbalanced training samples across a large range of age classes.

III. Proposed Approach

In this section, we describe the proposed approach including label-sensitive deep metric learning (LSDML) and the extension of multi-source LSDML (M-LSDML) in details, respectively. Moreover, we present the differences with some related works compared with our proposed approach.

A. LSDML

To learn robust and discriminative feature similarity for facial age estimation, the basic idea of our LSDML is to exploit the *label correlation* among face samples in the transformed subspace. Unlike recent deep metric learning methods [37], [38] which utilize hand-crafted features to be fed to the deep networks, our model jointly optimizes both tasks of learning similarity and embedding features for face representation in a unified deep architecture. Let $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote the training set which consists of N samples, where $\mathbf{x}_i \in \mathbb{R}^D$ denote the ith face image of D pixels and $y_i \in \mathbb{R}^1$ is the groundtruth age value, respectively. Our model is to compare the distance of face pairs by computing the feature representation $f(\mathbf{x}_i)$ for the ith face image $\mathbf{x}_i$ via deep neural networks. In terms of network architecture, we employ the residual learning method to optimize the whole network parameters, which have achieved superior performance in a volume of visual recognition tasks [43]. To better measure the learned face descriptors, we apply L_2 normalization on the obtained outcomes from the fully connected layers.

By feeding the face data to the designed network, we compute the similarity for each face pair of $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ as follows:

$$d_f(\mathbf{x}_i, \mathbf{x}_j) = \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2, \tag{1}$$

where $\|\cdot\|_2$ denotes the Euclidean distance beyond the learned embedded metric parameterized by $f(\cdot)$. The deep feature embedding for face representation is computed as:

$$f(\mathbf{x}_i) = \text{pool}\left(\text{ReLU}(\mathbf{W} \otimes \mathbf{x}_i + \mathbf{b})\right), \tag{2}$$

where $\otimes$ denotes the convolution operation, pool$(\cdot)$ denotes the max pooling operation, an ReLU$(\cdot)$ denotes the rectifier nonlinear function.

The crucial part of our LSDML is to learn the network parameters $f(\cdot)$. To achieve this goal, we first pass a given mini-batch forward the deep network, and we select each quadruplet of (i, j, k, l), such that $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}$, $(\mathbf{x}_i, \mathbf{x}_k) \in \mathcal{N}$ and $(\mathbf{x}_j, \mathbf{x}_l) \in \mathcal{N}$, where $\mathcal{P}$ and $\mathcal{N}$ denote the positive and negative pair set, respectively. More details are illustrated in Fig. 3. Moreover, to achieve the discriminativeness of the feature similarity, our LSDML enforces each $d_f(\mathbf{x}_i, \mathbf{x}_j)$ pair in positive set is close to each other, and at the same time $d_f(\mathbf{x}_i, \mathbf{x}_j)$ and $d_f(\mathbf{x}_k, \mathbf{x}_l)$ in negative set is pushed far away. As a result, the distance of inter-class pairs is minimized, and the distance of intra-class pairs is larger than a margin τ in the transformed subspace. To better measure the age difference information, we design a smoothing function $C(\cdot, \cdot)$ to measure the degree of any two age labels as follows:

$$C\left(y_i, y_j\right) = \exp^{\frac{-(y_i - y_j)^2}{H^2}}, \tag{3}$$

where H denotes the label difference threshold to determine the variance of age label distribution.

For the computational efficiency for sampling during training process, our LSDML automatically learns to select hard meaningful examples, so that the local manifold of face samples is preserved in the transformed subspace. Specifically, we first feed the mini-batched data to the designed deep network and then compute the feature descriptors for face representations. Having obtained these features, we select a hard quadruplet of $(\hat{i}, \hat{j}, \hat{k}, \hat{l})$, where the positive pair $d_f(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{j}})$

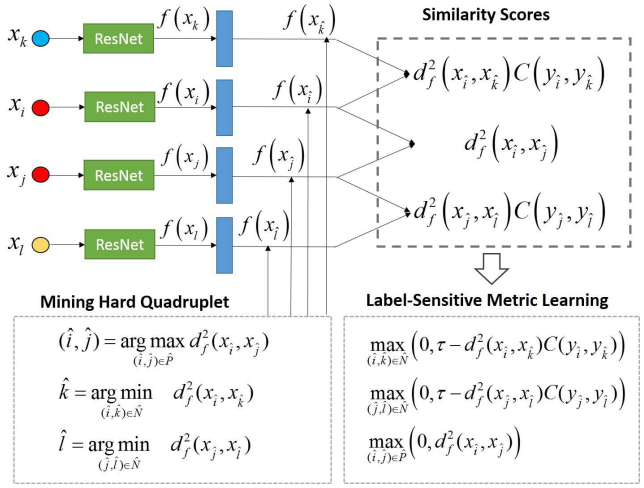


Fig. 3. The detailed learning procedure of our LSDML. First, our LSDML is fed with a mini-batch to the designed deep residual network. Then, to involve with both tasks of label-sensitive metric learning and hard negatives mining, our model aims to optimize: 1) intra-class separability and inter-class compactness; 2) isotonic similarity to the age difference degree, and 3) mining quadruplets of hard negatives. In this way, our model learns to mine hard negatives to preserve the local similarity of the data samples, so that the label correlation for ages is exploited in the embedded metric space. Note that the parameters of all ResNet-101 architectures are shared and optimized via the standard back-propagation method.

is with a largest similarity score, and both the negative pairs $d_f(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{k}})$ and $d_f(\mathbf{x}_{\hat{j}}, \mathbf{x}_{\hat{l}})$ are with smaller similarity scores during training process. Then our model focuses on optimizing these selected hard examples and the network parameters are learned via the back-propagation process. As a result, both tasks of learning feature similarity and mining meaningful hard examples are addressed simultaneously in a unified framework. We will present the formulation and optimization procedure in the following.

We formulate the objectives of our LSDML, which aims to minimize the following optimization problem:

$$\begin{aligned}
\min_{f} \; & J = J_1 + \lambda J_2 + \mu J_3 \\
& = \sum_{(\hat{i},\hat{j},\hat{k},\hat{l})} \left(\varepsilon_{\hat{i},\hat{k}} + \epsilon_{\hat{j},\hat{l}}\right) + \lambda \sum_{(\hat{i},\hat{j})} \rho_{\hat{i},\hat{j}} + \mu \|\mathbf{W}\|_F^2, \\
\text{subject to} \; & \max_{(\hat{i},\hat{k})\in\hat{\mathcal{N}}} \left(0, \tau - d_f(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{k}}) C(y_{\hat{i}}, y_{\hat{k}})\right)^2 \le \varepsilon_{\hat{i},\hat{k}}, \\
& \max_{(\hat{j},\hat{l})\in\hat{\mathcal{N}}} \left(0, \tau - d_f(\mathbf{x}_{\hat{j}}, \mathbf{x}_{\hat{l}}) C(y_{\hat{j}}, y_{\hat{l}})\right)^2 \le \epsilon_{\hat{j},\hat{l}}, \\
& \max_{(\hat{i},\hat{j})\in\hat{\mathcal{P}}} \left(0, d_f(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{j}})\right)^2 \le \rho_{\hat{i},\hat{j}}, \\
& \varepsilon_{\hat{i},\hat{k}} \ge 0, \quad \epsilon_{\hat{j},\hat{l}} \ge 0, \quad \rho_{\hat{i},\hat{j}} \ge 0,
\end{aligned} \tag{4}$$

where $\varepsilon_{\hat{i},\hat{k}}, \epsilon_{\hat{j},\hat{l}}, \rho_{\hat{i},\hat{j}}$ denotes the latent variables, τ denotes the thresholding margin (assigned to 1 in our experiments) and the hard quadruplet $(\hat{i}, \hat{j}, \hat{k}, \hat{l})$ is selected by following violated criterions:

$$\begin{aligned}
(\hat{i}, \hat{j}) &= \arg\max_{(\hat{i},\hat{j})\in\hat{\mathcal{P}}} \; d_f^2(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{j}}), \\
\hat{k} &= \arg\min_{(\hat{i},\hat{k})\in\hat{\mathcal{N}}} \; d_f^2(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{k}}), \\
\hat{l} &= \arg\min_{(\hat{j},\hat{l})\in\hat{\mathcal{N}}} \; d_f^2(\mathbf{x}_{\hat{j}}, \mathbf{x}_{\hat{l}}),
\end{aligned}$$

where $f(\mathbf{x})$ denotes the deep feature embedding from raw face image, $f(\cdot)$ is the learned deep network parameterized by $\{\mathbf{W}, \mathbf{b}\}$, λ is used to balance the inter-class compactness, the intra-class separability, μ denotes the regularization term and $\|\mathbf{W}^{(m)}\|_F^2$ denotes the Frobenius norm of matrix $\mathbf{W}^{(m)}$ to prevent the network parameters from overfitting.

There are four objectives for (4):

1) J_1 in (4) ensures that the distance of negative face pairs is maximized larger than a threshold, while J_2 in (4) ensures the distance of positive face pairs is minimized. As a result, both the inter-class compactness and the intra-class separability is exploited simultaneously in the learned feature similarity.
2) The proposed measurement $C(\cdot,\cdot)$ in J_1 of age label degree is applied to smooth the negative face pairs with different age value gaps. Hence, the ranking information and correlation of age labels is embedded in the transformed subspace.
3) For each mini-batched data, the hard samples are mined during the network optimization process. Moreover, the violations of training samples are optimized and back-propagated to the previous layers, so that the manifold of face samples is preserved locally in the learned subspace.
4) The joint terms jointly optimize both tasks of learning feature similarity and mining hard examples in a ranking-preserving manner. This enables us to jointly optimize the two that benefit each other, and the hard example-aware sampling method improves the computational efficiency during training process.

To optimize (4), we leverage the standard back-propagation method to update the parameters literately. To achieve this, we reformulate our objective function as follows:

$$\begin{aligned}
\min_{f} \; & J = J_1 + \lambda J_2 + \mu J_3 \\
& = \sum_{(\hat{i},\hat{j},\hat{k},\hat{l})} [\max_{(\hat{i},\hat{k})\in\hat{\mathcal{N}}} \left(0, \tau - d_f(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{k}}) C(y_{\hat{i}}, y_{\hat{k}})\right)^2 \\
& \quad + \max_{(\hat{j},\hat{l})\in\hat{\mathcal{N}}} \left(0, \tau - d_f(\mathbf{x}_{\hat{j}}, \mathbf{x}_{\hat{l}}) C(y_{\hat{j}}, y_{\hat{l}})\right)^2 \\
& \quad + \lambda \max_{(\hat{i},\hat{j})\in\hat{\mathcal{P}}} \left(0, d_f(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{j}})\right)^2] + \mu \|\mathbf{W}\|_F^2.
\end{aligned} \tag{5}$$

Since each learned face descriptor $f(x)$ is L_2 normalized, we can conveniently derive the gradients for the face distance $d_f^2(\mathbf{x}_i, \mathbf{x}_j)$. Note that the optimization mainly depends on violated quadruplets which cannot satisfy the expression evaluates in $\max(\cdot)$ to true and outputs 0 otherwise. Having obtained the gradients, the network parameters $\mathbf{W}$ and $\mathbf{b}$ are updated by using the gradient-decent algorithm as follows until convergence:

$$\mathbf{W} = \mathbf{W} - \eta \frac{\partial J}{\partial \mathbf{W}}, \tag{6}$$

$$\mathbf{b} = \mathbf{b} - \eta \frac{\partial J}{\partial \mathbf{b}}, \tag{7}$$

where η is the learning rate, which controls the convergence speed of the objective function J.

Algorithm 1 shows the optimization procedure of LSDML.

Algorithm 1 LSDML

Input: Training Set: $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, $H = 5$, $\lambda = 0.4$ and $\mu = 0.001$, η.

Output: Network Parameters $\{\mathbf{W}, \mathbf{b}\}$.

Step 1 (Parameters Intialization): Initialize the parameters $\{\mathbf{W}, \mathbf{b}\}$ by pretrained models such as ResNet-101 [44].

Step 2 (Optimization via Back-Propagation):

repeat

2.1 Passing a mini-batched data forward the designed deep neural networks (*i.e.*, ResNet-101), and select hard quadruplet $(\hat{i}, \hat{j}, \hat{k}, \hat{l})$.

2.2 Performing backward propagation and compute the gradients.

for $(\hat{i}, \hat{j}) \in \hat{\mathcal{P}}$ **do**

$$\frac{\partial J}{\partial f(\mathbf{x}_{\hat{i}})} \leftarrow \frac{\partial J}{\partial d_f^2(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{j}})} \cdot \frac{\partial d_f^2(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{j}})}{\partial f(\mathbf{x}_{\hat{i}})}$$

$$\frac{\partial J}{\partial f(\mathbf{x}_{\hat{j}})} \leftarrow \frac{\partial J}{\partial d_f^2(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{j}})} \cdot \frac{\partial d_f^2(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{j}})}{\partial f(\mathbf{x}_{\hat{j}})}$$

for $(\hat{i}, \hat{k}) \in \hat{\mathcal{N}}$ **do**

$$\frac{\partial J}{\partial f(\mathbf{x}_{\hat{i}})} \leftarrow \frac{\partial J}{\partial d_f^2(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{k}})} \cdot \frac{\partial d_f^2(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{k}})}{\partial f(\mathbf{x}_{\hat{i}})}$$

$$\frac{\partial J}{\partial f(\mathbf{x}_{\hat{k}})} \leftarrow \frac{\partial J}{\partial d_f^2(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{k}})} \cdot \frac{\partial d_f^2(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{k}})}{\partial f(\mathbf{x}_{\hat{k}})}$$

for $(\hat{j}, \hat{l}) \in \hat{\mathcal{N}}$ **do**

$$\frac{\partial J}{\partial f(\mathbf{x}_{\hat{j}})} \leftarrow \frac{\partial J}{\partial d_f^2(\mathbf{x}_{\hat{j}}, \mathbf{x}_{\hat{l}})} \cdot \frac{\partial d_f^2(\mathbf{x}_{\hat{j}}, \mathbf{x}_{\hat{l}})}{\partial f(\mathbf{x}_{\hat{j}})}$$

$$\frac{\partial J}{\partial f(\mathbf{x}_{\hat{l}})} \leftarrow \frac{\partial J}{\partial d_f^2(\mathbf{x}_{\hat{j}}, \mathbf{x}_{\hat{l}})} \cdot \frac{\partial d_f^2(\mathbf{x}_{\hat{j}}, \mathbf{x}_{\hat{l}})}{\partial f(\mathbf{x}_{\hat{l}})}$$

2.3 Updating the parameters according to (6) and (7).

until *convergence or reaching the maximum iteration* T;

Return: $\{\mathbf{W}, \mathbf{b}\}$.

B. Multi-Source LSDML

In this subsection, we extend our LSDML to a multi-source LSDML (M-LSDML) framework for facial age estimation. Since it is difficult to densely collect face samples which cover a large range of age labels in the real-world applications, there exists few labeled samples belong to some age labels, which may bias data distribution for certain age classes. To address this problem, grouping age labels techniques [4], [45] have been proposed to split age pattern by a series of age groups and exploit the smoothness between the neighbouring groups. However, these methods are hand-crafted, which may destroy the real-world aging pattern in the cross-population datasets. To circumvent this problem, our M-LSDML aims to learn the label-sensitive feature similarity by including more than one source face aging dataset that is available for training. To achieve this, our proposed M-LSDML is to maximize the correlation of different source datasets as illustrated in Fig. 4, which reinforces the robustness and discriminativeness of the learned feature similarity by exploiting the label correlation.

Suppose we have M_s cross-population training datasets, our model aims to maximize the correlation of the multi-source datasets as targets as well as among all source population.

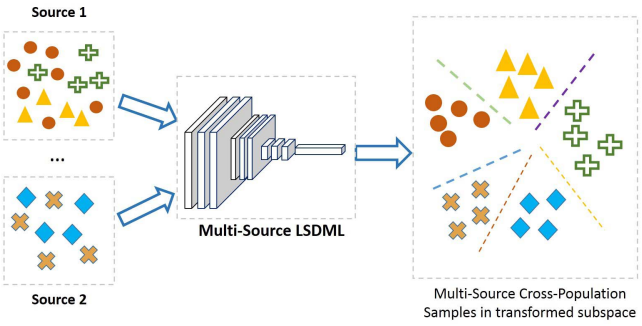


Fig. 4. The illustration of the proposed multi-source LSDML framework, which aims to learn a label-sensitive and discriminative feature similarity by utilizing multi-source datasets. Specifically, our multi-source LSDML aims to maximize the correlation of different datasets via deep network, which reinforces the discriminativeness of the feature similarity in the transformed subspace.

Hence, we formulate the objective function as follows:

$$\begin{aligned}
\min_f J &= J_1 + \lambda J_2 + \mu J_3 \\
&= \sum_{s=1}^{M_s} \sum_{t=s+1}^{M_s} \left[\sum_{(\hat{i},\hat{j},\hat{k},\hat{l})} \left(\varepsilon_{\hat{i},\hat{k}} + \epsilon_{\hat{j},\hat{l}} \right) + \lambda \sum_{(\hat{i},\hat{j})} \rho_{\hat{i},\hat{j}} \right] + \mu \|\mathbf{W}\|_F^2, \\
\text{subject to } & (\hat{i}, \hat{j}, \hat{k}, \hat{l}) \in \mathcal{N}_s \cup \mathcal{N}_t, \quad (\hat{i}, \hat{j}) \in \mathcal{P}_s \cup \mathcal{P}_t, \\
& \max_{(\hat{i},\hat{k}) \in \hat{\mathcal{N}}} \left(0, \tau - d_f(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{k}}) C(y_{\hat{i}}, y_{\hat{k}})\right)^2 \le \varepsilon_{\hat{i},\hat{k}}, \\
& \max_{(\hat{j},\hat{l}) \in \hat{\mathcal{N}}} \left(0, \tau - d_f(\mathbf{x}_{\hat{j}}, \mathbf{x}_{\hat{l}}) C(y_{\hat{j}}, y_{\hat{l}})\right)^2 \le \epsilon_{\hat{j},\hat{l}}, \\
& \max_{(\hat{i},\hat{j}) \in \hat{\mathcal{P}}} (0, d_f(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{j}}))^2 \le \rho_{\hat{i},\hat{j}}, \\
& \varepsilon_{\hat{i},\hat{k}} \ge 0, \quad \epsilon_{\hat{j},\hat{l}} \ge 0, \quad \rho_{\hat{i},\hat{j}} \ge 0, \\
& (\hat{i}, \hat{j}) = \arg\max_{(\hat{i},\hat{j}) \in \hat{\mathcal{P}}} d_f^2(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{j}}), \\
& \hat{k} = \arg\min_{(\hat{i},\hat{k}) \in \hat{\mathcal{N}}} d_f^2(\mathbf{x}_{\hat{i}}, \mathbf{x}_{\hat{k}}), \\
& \hat{l} = \arg\min_{(\hat{j},\hat{l}) \in \hat{\mathcal{N}}} d_f^2(\mathbf{x}_{\hat{j}}, \mathbf{x}_{\hat{l}}),
\end{aligned} \tag{8}$$

where $\beta (\beta > 0)$ denotes a hyper-parameter to balance these terms, and the sets of $\mathcal{N}_s$, $\mathcal{N}_t$, $\mathcal{P}_s$, $\mathcal{P}_t$ denote the face pair sets which are drawn from the union of any two multi-source indexed by the sth and tth datasets.

To solve the optimization problem in (8), we leverage the stochastic gradient-decent method to compute the gradients and update the parameters. With the learned label-sensitive similarity, the label correlation is smoothly preserved in the learned transformed subspace thanks to the discriminative knowledge captured from multi-source cross-population datasets. Moreover, the bias of training samples is efficiently removed by maximizing the correlation of multi-source datasets. We will show the effectiveness of our M-LSDML on multi-source facial age estimation datasets in our experiments.

IV. Experiments

In this section, we evaluated the performance of the proposed approach on four standard face aging datasets including the MORPH (Album2) [46], FG-NET [22], FACES [47],

AdienceFaces [48] and ChaLearn [49] datasets. In particular, the face samples in both ChaLearn [49] and AdienceFaces [48] datasets were captured in wild conditions, so that they undergo various challenging situation such as different facial expressions, aspect ratios, clutter background, scale variations, in-plane rotation and diverse partial occlusions.

A. Datasets

1) MORPH (Album 2) [46]: This dataset consists of about 55000 face images ranging from 16 to 77 years old, which is a large scale publicly dataset annotated with accurate age values. There are averaging 3 samples per person, and the age gap is averagely 1.35±1.65 years per person. Hence, only a short-term growth process is gathered for each individual. There are many ethnicity (White, Black, Hispanic, Asian, and others) in this dataset.

2) AdienceFaces [48]: The dataset consists of 26,580 face images belonging to 2,284 subjects, where these samples were manually labeled by the age groups (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60-100) and a subset of age values such as 3, 35, 55, etc. All images were captured in the wild conditions, which undergoes the variations of appearances, noises, large poses and partial occlusions.

3) FG-NET [22]: This dataset consists of 1002 images with 82 persons and there exists averaging 12 samples for each person. The age range covers from 0 to 69. For every person, the age gap is averagely 27.80±11.75 years, reflecting a relatively long-term growth detail of the person. The face samples in the dataset encounter large variations in aspect ratios, illumination and diverse expressions.

4) FACES [47]: This dataset contains 2052 face images from 171 persons. The age range covers from 19 to 80 years old. For each person, there are six expressions including neutral, sad, disgust, fear, angry and happy. The large variances of diverse facial expressions bring label ambiguity for age predicting.

5) ChaLearn [49]: This dataset is drawn from the apparent age estimation challenge [49], which contains 5000 images for training and 1500 images for validation. The age range covers from 0 to 100 years old, which were collected from social networks. The face images suffer from large variations of diverse facial expressions, poses and partial occlusions. Each face in the dataset was labeled by ten persons, and the mean apparent age and the standard deviation were used for annotation.

B. Experimental Settings

Before evaluation, we detected facial bounding boxes for each given facial image via the DLIB [50] image processing library. Then we resized the detected faces to the size of 224×224 that matches the input dimension D of the employed deep networks. For each facial image, we detected three landmarks including two centers of eyes and the nose base to align the face into the canonical coordinate system by using affine transformation. Since our approach aims to learn a transformed subspace to measure the feature similarity of face pairs, we passed the given facial image to the learned network and compute the descriptors for face representation. Having obtained these face features, we trained an age estimator by OHRANK [9] and then obtained exact age values for evaluation. It should be notified that we trained the OHRANK by utilizing only single-source dataset under the M-LSDML evaluation setting for fair experimental comparisons.

C. Implementation Details

The proposed methods were implemented under the open source CAFFE [51] deep learning toolbox, which has been widely used for the deploying and evaluation for deep architectures. For the hyper-parameters employed in our proposed LSDML, we set $H = 5$, $\lambda = 0.4$ and $\mu = 0.001$ by cross-validation. The residual network [44] consists of a series of small 3×3 receptive fields (convolution layers), pooling operations and the nonlinear ReLU rectifier function. In particular, the residual network was equipped with a skip-connection (identity module), which memorized the residual estimation and avoid decent vanishing though the very deep architecture. Note that we flattened the feature maps of last layer and then adopted one-layer fully connections to reduce them to 50-dimension feature vector. For the hyper-parameters of the network, we specified the values of the weight decay, moment empirically to 0.0001, 0.9, respectively. The learning rates in the training stage were tuned and decreased over the interval of every 1000 epochs. The whole training procedure converged until the validation error remained minimized and unchanged. Lastly, we randomly oversampled all face images during training process by horizontal flipping and shuffling to generate more training samples to reinforce the network generation. It is important to initialize the networks parameters $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$, where m denotes the layer number of the deep networks. In our experiments, we initialized the parameters of the remaining layers by utilizing the pretrained deep networks, and leveraged the uniform distribution [52] to initialize $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$.

D. Evaluation Protocols

1) Mean Absolute Error: For the evaluation metrics, we utilized the mean absolute error (MAE) [1], [19], [25], [33] to measure the error between the predicted age and the ground-truth, which is computed as follows:

$$e = \frac{\sum_{i=1}^{N} \left\| \hat{y}_i - y_i^* \right\|_2}{N} \tag{9}$$

where $\hat{y}$ and y^* denote predicted and ground-truth age value, respectively, and N denotes the number of the testing samples.

2) Cumulative Score Curve: We also applied the cumulative score (CS) [23], [24], [26], [33] curve to quantitatively evaluate the performance of age estimation methods. The cumulative prediction accuracy at the error ϵ is computed as:

$$CS(n) = \frac{K_n}{K} \times 100\% \tag{10}$$

where K is the total number of testing images, K_n is the number of testing images whose absolute error between the estimated age and the ground-truth age is not greater than n years.

TABLE I
COMPARISON OF MAEs OF OUR APPROACH WITH DIFFERENT STATE-OF-THE-ART APPROACHES ON THE MORPH DATASET

Methods	Model Description	MAE	Year
BIF+KNN	-	9.64	-
AGES [1]	AAM + facial aging pattern	8.83	-
MTWGP [23]	AAM + Multi-task warped Gaussian progress regression	6.28	2010
BIF+OLPP [53]	Age estimation cross age and gender	4.20	2010
CS-LDA [54]	Cost-sensitive subspace learning	6.03	2010
Raw+OHRANK [9]	Paw pixel + Ordinal Hyperplane Ranking	7.34	2011
LBP+OHRANK [9]	LBP + Ordinal Hyperplane Ranking	6.88	2011
BIF+OHRANK [9]	BIF + Ordinal Hyperplane Ranking	6.49	2011
CS-FS [55]	Cost-sensitive feature selection	6.59	2012
IIS-LDL [12]	AAM/BIF + label distribution learning	5.69	2013
CPNN [12]	AAM/BIF + label distribution learning	5.67	2013
CA-SVR [56]	AAM+cumulative /joint attribute learning	4.87	2013
MFOR [57]	Multi-feature ordinal ranking	5.88	2013
rKCCA [58]	Multi-model canonical correlation analysis	3.98	2014
rKCCA + SVM [58]	Multi-model canonical correlation analysis + SVM classifier	3.91	2014
CS-LBFL [15]	Cost-sensitive local binary feature learning	4.52	2015
CS-LBMFL [15]	Cost-sensitive local binary multiple feature learning	4.37	2015
CS-OHR [59]	Cost-sensitive ordinal hyperplane ranking	3.74	2015
DeepRank [33]	Scattering transforms + ordinal hyperplane ranking	3.57	2015
DeepRank+ [33]	Scatter transforms + ordinal hyperplane ranking	3.49	2015
OR-CNN [19]	Ordinal regression with deep convolutional networks	3.27	2016
LSDML	Label-sensitive deep metric leaning	**3.08**	-
M-LSDML	Multi-source label-sensitive deep metric leaning	**2.89**	-

E. Evaluation on MORPH (Album2)

In our settings, we performed five folds cross-validation of our proposed approach on the MORPH (Album2) dataset. Specifically, we divided the whole dataset into five equal-size folds. Then we used one fold (20% of total data) as the testing set and the other four folds (80% of total data) as the training set. We repeated this procedure ten times and finally averaged the results as the facial age estimation results.

1) Comparisons With the State-of-the-Art Methods: We trained our LSDML with the training set of MORPH (Album 2). Table I tabulates the MAEs and Fig. 5 shows the CS curves of our LSDML in comparisons to the state-of-the-art age estimation methods with the standard evaluation protocol, respectively. From these results, we see that our LSDML achieves higher age estimation performance than the state-of-the-art features such as raw pixels, local binary patter (LBP) [7] and bio-inspired feature (BIF) [11]. This is because our LSDML learns discriminative face descriptors directly from raw pixels, which achieves label correlation for ages and robustness to large variations of face images in unconstrained environments. Compared with the feature learning-based methods such as CS-LBFL [15] and CS-LBMFL [15], our LSDML demonstrates better performance, which achieves the nonlinear relationship between the face samples and consecutive age labels. We also compared our proposed LSDML with the deep learning-based methods including DeepRank [33], DeepRank+ [33] and OR-CNN [19] that utilized deep neural networks. As results shown in Table I and Fig. 5, we see that our method consistently outperforms those deep models on this dataset, which shows the effectiveness of the proposed LSDML, where the complementary information of label-sensitive similarity learning and hard examples mining is exploited to reinforce our model.

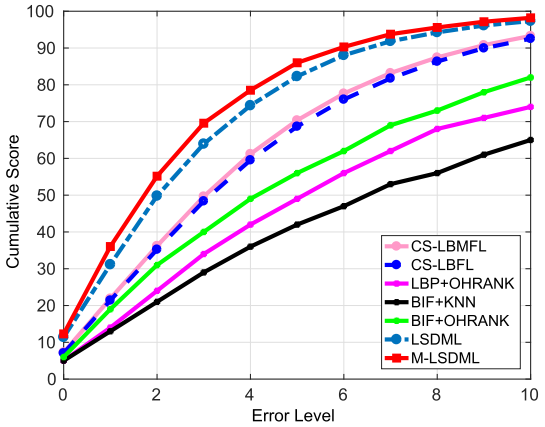


Fig. 5. The CS curves of our approach compared with different facial age estimation methods on the MORPH dataset.

Moreover, we evaluated the proposed M-LSDML by utilizing the MORPH (Album2), FG-NET, ChaLearn, FACES as the multi-source training face data. Specifically, we trained our M-LSDML model by leveraging MORPH (Album2) training samples, FG-NET, ChaLearn, FACES and LIESPAN datasets, and the MORPH testing data were used for evaluation. As the results are shown in Table I and Fig. 5, we find out that M-LSDML outperforms LSDML by about 0.25 years old in terms of the MAE performance. This is because our M-LSDML maximizes the correlation of the auxiliary cross-population datasets in order to learn more robust and discriminative feature similarity, while our LSDML trained in single-source dataset that biases data distribution across a large range of age classes.

TABLE II
COMPARISON OF MAEs OF OUR APPROACH WITH DIFFERENT DEEP METRIC LEARNING METHODS ON THE MORPH DATASET

Methods	MAE	CED$_{\theta\leq 1}$	CED$_{\theta\leq 5}$
Contrastive Loss [36]	3.72	20.5%	67.1%
Triplet Hinge Loss [37]	3.59	24.6%	73.4%
Lifted Structural Loss [39]	3.24	30.3%	81.0%
LSDML	**3.08**	**32.7%**	**82.9%**
M-LSDML	**2.89**	**38.2%**	**86.1%**

TABLE III
COMPARISON OF MAEs OF OUR METHOD COMPARED WITH DIFFERENT DEEP NETWORKS ARCHITECTURES ON THE MORPH DATASET

Methods	Network Architectures	MAE
SqueezeNet	Compressed Network [60]	3.89
AlexNet	AlexNet [61]	3.72
GoogleNet	Inception V3 [62]	3.49
LightenCNN	Squeezed Maxout for Face [63]	3.97
VGG-16	VGG Face Net [64]	2.91
ResNet-101	ResNet-101 for Face [44]	**2.89**

TABLE IV
COMPARISON OF MAEs OF OUR APPROACH WITH DIFFERENT AGE ESTIMATORS ON THE MORPH DATASET

Methods	MAE
ResNet-101 [44] + KNN	4.71
ResNet-101 [44] + Single Label	3.55
ResNet-101 [44] + Gaussian Label	3.29
ResNet-101 [44] + OR-CNN [19]	3.33
ResNet-101 [44] + LSDML	**3.08**
ResNet-101 [44] + M-LSDML	**2.89**

2) Comparisons With Different Deep Metric Learning Methods: To demonstrate the effectiveness of the learned metric, we conducted experimental comparisons with different deep metric learning methods including contrastive loss [36], triplet loss [41] and lifted structural loss [39]. To fairly compare our LSDML model with previous metric learning methods, we directly deployed the loss functions at the top of ResNet-101 and fine-tuned the network parameters via the back-propagation method. Table II qualifies the MAE and CED curves performance of our LSDML in comparisons to the other deep metric learning methods. According to these results, we see that our LSDML consistently outperforms the state-of-the-art deep metric learning methods which also utilize the large margin constraints on the similarity of face pairs. The main reason is that our LSDML achieves age difference information in the learned metric space, while the state-of-the-arts consider the feature similarity equivalently which ignores the age correlation and performs worse than our LSDML. In particular, compared with the quadruplet-based lifted structural technique, our LSDML performs better performance, because our model automatically achieves useful and age-adaptive information in the embedded subspace.

3) Comparisons With Different Network Architectures: We investigated the performance effects of tuning different network architectures with our LSDML. Specifically, we compared our model with existing deep networks including SqueezeNet [60], AlexNet [61], GoogleNet [62], LightenCNN [63] VGG-16 Face Net [64] and ResNet-101 Face Net [44]. In particular, the SqueezeNet, AlexNet and GoogleNet were pretrained by a large scale of ImageNet [61] images, while the LightenCNN, VGG-16 and ResNet-101 were pretrained by face images. Before evaluation, we rescaled the raw inputs as the required dimension of each network. To be specific, we resized the face image as the size of 227×227 for AlexNet, 128×128 of gray images for SqueezeNet and LigntenCNN, and 224×224 for VGG-16 and ResNet-101 networks. We discarded the last classification layer and fine-tuned these mentioned networks by our proposed loss function. Note that we updated the total parameters through the whole network during the back-propagation process. Table III tabulates the results of performance effects with different deep networks based on our method. From these results, we have gained two observations: 1) the very deep network VGG-16 and ResNet-101 pretrained by face images outperforms those pretrained by ImageNet data, and 2) very deep face networks including VGG-16 and ResNet-101 architectures significantly improve the facial age estimation performance. It is valuable to note that ResNet-101 trained by augmented face samples by a 3D morphable face model, which obtains very outperformed performance compared with VGG-16 face model pretrained by millions of face data.

4) Comparisons With Different Age Estimators: We investigated the performance effects of our LSDML with different facial age estimators. To fair comparisons, we firstly created the baseline method by ResNet-101 [44] and KNN. It is notified that we directly passed face images forward to ResNet and obtained the face representation. Then, we deployed the softmax loss [61] as the single label method, and the deep label distribution learning [65] as the Gaussian label methods at the top of ResNet-101 [44] and finetuned these networks. Moreover, we compared our model with the state-of-the-arts facial age estimation method OR-CNN [19] which leverages deep convolutional networks to learn age-informative feature representation. As the results are showed in Table IV, we see that our model achieves better performance in comparisons with the other baselines. Even our method outperforms OR-CNN [19], which demonstrates the effectiveness of jointly optimizing both tasks of our label-sensitive metric learning and hard example mining in order to benefit the complementary information from each other.

5) Computational Time: Lastly, we compared the computational time of our approach which cooperates with quadruplet-based hard example mining. First, we created hard quadruplet mining based on Euclidean distance space as the baseline method. As shown in Table V, we see that our proposed method achieves about 1.5× faster convergence speed compared with the baseline method. Moreover, our LSDML copes comfortably with the violated hard samples, so that the local manifold for the input data points is preserved in the learned metric space. These results also demonstrate the advantages of the proposed sampling strategy in our LSDML, which aims to optimize the hard meaningful examples on the update process of the network parameters during back-propagation process.

TABLE V

COMPUTATIONAL TIME OF OUR APPROACH WITH OUR PROPOSED QUADRUPLET+LSDML VERSUS QUADRUPLET+EUCLIDEAN ON TRAINING PROCESS ON THE MORPH DATASET

Methods	Training Time	Epoches
Quadruplet+Euclidean [39]	∼30 h	12000
Quadruplet+LSDML	∼24 h	8000

TABLE VI

COMPUTATION TIME OF OUR METHOD COMPARED WITH DIFFERENT NETWORK ARCHITECTURES IN THE MORPH DATASET UNDER THE GPU WITH NVIDIA GTX 1080. NOTE THAT DFD, LQP, RICA AND CS-LBFL WERE TESTED ON CPU PLATFORM

Methods	Testing Time (imgs/s)
DFD [66]	2
LQP [67]	10
RICA [68]	3.5
CS-LBFL [15]	20
SqueezeNet [60]	3582.6
AlexNet [61]	2425.3
LightenCNN [63]	2173.2
GoogleNet [62]	346.2
ResNet-101 [44]	256.8
VGG-16 [64]	143.2

We also investigated the computational time with different deep networks based on the efficient Caffe [51] toolbox, and the whole architectures were built on a speed-up parallel computing GPU with NVIDIA GTX 1080. Table VI tabulates the comparisons of the computational time during the testing phase. From these results, we see that the deep architectures achieve the real-time age estimation with a GPU for the feature extraction procedure. In addition, we carefully implemented the OHRANK by following the details provided in [9]. The OHRANK takes 0.04 seconds by using an Intel i5-CPU@3.20GHz PC, which also satisfies the real-time requirements.

F. Evaluation on AdienceFaces

1) Comparisons With the State-of-the-Arts Methods: For evaluation, we utilized the aligned faces pre-processed by the provided alignment tool[1] and leveraged the five-fold cross validation on these face data. We created a deep learning based approach with the VGG Face Net [64] and the softmax classification loss function as the baseline. By carefully following the evaluation protocol employed in [32], we used the classification accuracy and 1-off accuracy which means the predicted label is within the neighbouring groups of the true one for comparisons. It is worthwhile to note that we trained our LSDML by the AdienceFaces [48] training set, while for our M-LSDML we leveraged the unconstrained age grouping datasets including the AdienceFaces [48] training set, MORPH [46] and the ChaLearn [49] training and validation sets (we grouped face samples in ChaLearn [49] the same procedure as [31]). Table VII tabulates the results of our methods compared with [32] on the AdienceFaces [32] dataset. According to the results, we observe that the proposed methods obtains better performance than Cascaded-CNN [32] especially on the 1-off accuracy. Hence, with the proposed label-sensitive metric and deep architecture, our models achieve the promising results on this dataset where the face samples undergo large variances of poses, lightness and cluttered backgrounds.

[1] http://www.openu.ac.il/home/hassner/Adience/code.html#inplanealign

TABLE VII

COMPARISONS OF MAES OF OUR APPROACH WITH THE STATE-OF-THE-ART METHODS ON THE ADIENCEFACES [48] DATASET THAT WERE CAPTURED IN THE WILD CONDITIONS. NOTE THAT 1-OFF ACCURACY MEANS THE PREDICTED LABEL IS TRUE WITHIN THE NEAREST-NEIGHBOUR GROUPS FOR COMPARISONS, TYPICALLY

Methods	Exact	1-off
Best from [48]	45.1±2.6	79.5±1.4
Best from [29]	50.7±5.1	84.7±2.2
Cascaded-CNN [32]	52.9± 6	88.5 ±2.2
VGG Face Net [64] + softmax	54.8±10.2	89.3±6.5
LSDML	**56.9±6.0**	**91.8±3.1**
M-LSDML	**60.2±5.3**	**93.7±2.3**

2) Performance Effects of Different Data Augmentation Strategies: To evaluate the effectiveness of our proposed M-LSDML in comparisons to conventional data augmentation methods, we conducted experiments of our methods compared with the horizontal flipping and random cropping. Specifically, the horizontal flipping mirrors the face samples based on the horizontal axis, while the random cropping randomly crops images from the original input by the specified translations during training (centering cropping during testing), respectively. Table VIII shows the results of our models compared with the horizontal flipping and random cropping on the AdienceFaces [48] dataset. From these results, we have made three-fold conclusions as follows:

1) By comparing our LSDML which utilizes data augmentation techniques with the proposed M-LSDML, we observe that our M-LSDML significantly outperforms the baselines with data augmentation methods such as horizontal flipping and random cropping by about 3% accuracy. Hence, the results demonstrate that our M-LSDML consider large variances due to the cross-population face aging datasets.
2) The reporting results indicates that our M-LSDML significantly outperforms LSDML with multiple external facial age estimation datasets, which shows the effectiveness of the employed label correlation framework in M-LSDML. In addition, LSDML learned by external data exhibits the comparable performance with LSDML with data augmentation strategies including random cropping and horizontal flipping. This is because LSDML by simply combining datasets ignored to maximize the correlation for cross-population datasets, leading to scarce improvements for LSDML even with multi-source face aging data.
3) By comparing our model regarding with versus without any data augmentation methods like horizontal flipping and random cropping, we see that these data augmentation-based techniques slightly improve the performance for both our LSDML and M-LSDML.

TABLE VIII

COMPARISON OF CLASSIFICATION ACCURACY OF OUR M-LSDML WITH DIFFERENT DATA AUGMENTATION METHODS ON THE ADIENCEFACES [48] DATASET. NOTE THAT N-OFF MEANS THE PREDICTED LABEL IS TRUE WITHIN THE N-NEIGHBOUR GROUPS FOR COMPARISONS, TYPICALLY, THE VALUES OF N WERE SPECIFIED TO 1 AND 2

Methods	Augmentation Strategy	Exact	1-off	2-off	Training Dataset
LSDML	w/o data augmentation	56.0±8.2	91.7±2.7	97.3±0.9	D_1
LSDML	w/o data augmentation	56.8±6.9	90.3±1.5	98.0±1.4	D_1,D_2, D_3
LSDML	random cropping + horizontal flipping	56.9±6.0	91.8±3.1	97.9±0.9	D_1
LSDML	random cropping + horizontal flipping	57.3±7.2	92.4±4.0	97.6±1.1	D_1,D_2, D_3
M-LSDML	w/o data augmentation	58.2±7.5	93.3±2.9	98.6±1.3	D_1,D_2, D_3
M-LSDML	horizontal flipping	59.1±8.1	94.4±3.2	98.8±1.3	D_1,D_2, D_3
M-LSDML	random cropping	59.9±7.5	94.8±1.9	99.1±1.2	D_1,D_2, D_3
M-LSDML	random cropping + horizontal flipping	**60.2±5.3**	**93.7±2.3**	**98.2±0.7**	D_1,D_2, D_3

D_1-AdienceFaces [46] training set, D_2-MORPH [46], D_3-ChaLearn [49]

TABLE IX

COMPARISON OF MAES OF OUR APPROACH COMPARED WITH STATE-OF-THE-ART APPROACHES ON THE FG-NET DATASET

Methods	MAE	Year
BIF+KNN	8.24	-
RUN2 [69]	5.78	2007
AGES [1]	6.77	2007
LARR [26]	5.07	2008
PFA [70]	4.97	2008
MTWGP [23]	4.83	2010
RED-SVM [71]	5.21	2010
Raw+OHRANK [9]	6.25	2011
LBP+OHRANK [9]	4.92	2011
BIF+OHRANK [9]	4.48	2011
mKNN [72]	5.21	2012
LDL [12]	5.77	2013
CPNN [12]	4.76	2013
CA-SVR [56]	4.67	2013
CS-OHR [59]	4.70	2015
CS-LBFL [15]	4.43	2015
CS-LBMFL [15]	4.36	2015
LSDML	**3.92**	-
M-LSDML	**3.74**	-
Cascaded-CNN† [32]	3.49	2016
LSDML†	3.53	-
M-LSDML†	**3.31**	-

† By following the settings in [32], we randomly split the FG-NET [22] dataset into two folds: one fold consists of 890 samples for training and the remaining 112 samples were used for testing.

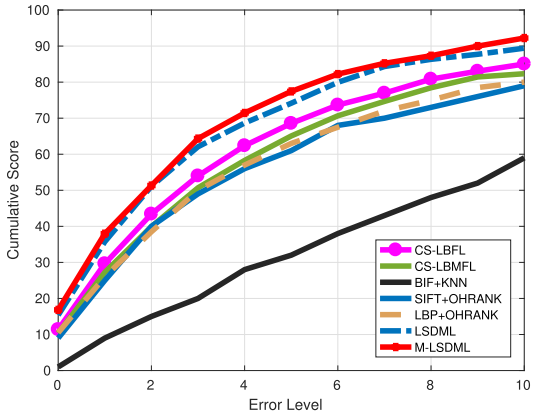


Fig. 6. The CS curves of our approach compared with different facial age estimation methods on the FG-NET dataset.

G. Evaluation on FG-NET

We evaluated our approach on the FG-NET dataset, which undergoes two challenges: 1) face samples encounter large aspect ratios and cluttered background, 2) there exits limited labeled training samples. Due to the limited samples for each subjected person, we adopted the leave-one-person-out (LOPO) strategy to conduct age estimation experiments. Specifically, we used all face images of one person as the test set and the remaining were used for training. We averaged the 82 folds results as the final age estimation performance. Since the 82-fold deeply training in ResNet-101 is time-consuming and easily leads to overfitting, in our experiments, we froze the previous convolution layers and only fine-tuned the last convolutional layer by feeding the provided face samples to the network during training process. Note that we set the value 0.0001 to the learning rate to fast convergence. We compared our LSDML trained by the training samples from FG-NET with the state-of-the-art facial age estimation evaluated in the FG-NET dataset in Table IX and Fig. 6. The demonstrated results show the effectiveness of our proposed LSDML, where the complementary information of both tasks of label-sensitive metric learning and hard example mining is exploited to make the learned metric more robust. We also evaluated the proposed M-LSDML on this dataset. Specifically, we chose the face samples for each person as the testing set and utilized the remaining FG-NET face samples by combining the union of the MORPH (Album2), ChaLearn, FACES datasets as the training set. Hence, the whole procedures also performed 82 folds and we averaged these results as the final performance. From these results, we see that the proposed M-LSDML improves the facial age estimation performance thanks to the full use of the correlation of multi-source face aging datasets, which achieves the more discriminative capacity for the learned feature similarity.

We also compared our methods with the facial age estimation method which was recently proposed in [32]. For fair comparisons with [32], we employed the 890-train and 112-test evaluation protocol by following the setting in [32]. We repeated the procedure for ten times and averaged the

TABLE X
COMPARISON OF MAEs OF OUR APPROACH WITH DIFFERENT STATE-OF-THE-ART APPROACHES ON THE FACES DATASET

Methods	Neutral	Happy	Disgust	Fearful	Sad	Angry	Year
BIF+OHRANK [9]	5.16	7.64	8.31	7.00	6.87	7.87	2011
LBP+OHRANK [9]	6.36	8.88	9.20	7.30	9.09	8.86	2011
BIF [73]	9.50	10.70	13.26	12.65	10.78	13.26	2012
BIF+MFA [73]	8.14	10.32	12.24	10.73	10.66	10.96	2012
CS-LBFL [15]	5.06	6.53	7.15	6.32	6.27	6.94	2015
CS-LBMFL [15]	4.84	5.85	5.70	6.10	4.98	5.50	2015
DeepRank [33]	5.99	7.12	8.15	6.35	7.77	6.68	2015
DeepRanker+ [33]	5.86	7.87	7.80	6.66	7.49	6.59	2015
LSDML	**3.88**	**3.49**	**4.41**	**5.10**	**4.09**	**3.87**	-
M-LSDML	**3.83**	**3.11**	**4.16**	**5.01**	**3.67**	**3.16**	-

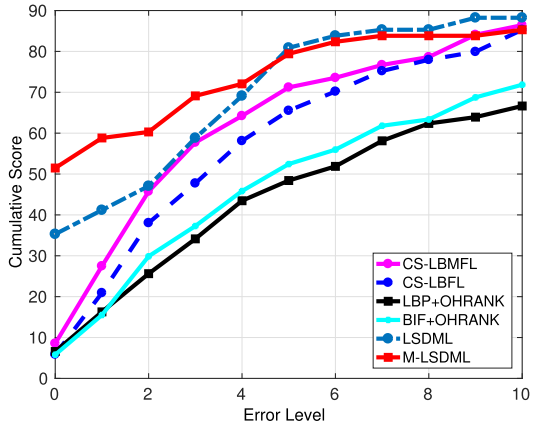


Fig. 7. The CS curves of our approach compared with different facial age estimation methods for Happy Expression on the FACES dataset.

results for final performance. Note that we trained our M-LSDML by utilizing the face aging datasets including FG-NET [22] training set (as noted 890 face samples), MORPH (Album2) [46], ChaLearn [49] and FACES [47] which were employed to predict the exact age values. According to these results, we observe that our methods outperforms Cascaded-CNN [32]. The achievements benefit from both the developed hard-mining strategy to discover the hard examples via the deep residual networks, which exploits the complex relationship of face images and age labels. However, one thing should be noted that evaluation on the FG-NET [22] by employing the LOPO protocol clarifies the experimental conclusions rather than an empirical performance evaluation, which has also been discussed in [32].

H. Evaluation on FACES

To demonstrate the advantages of our proposed approach, we evaluated our approach on the face aging dataset FACES which is exposed to diverse facial expressions. For fair comparisons, we conducted the experiments under the same expression. Note that our LSDML was trained by the training set from FACES, while our M-LSDML was learned by the MORPH, FG-NET datasets along with the training samples FACES. Fig. 7 shows the CS curves of our approach compared with different facial age estimation methods, and Table X tabulates the MAEs, respectively. According to these results, we see that our LSDML largely improves the facial age estimation performance compared with the state-of-the-arts, which shows the discriminativeness of the learned deep feature embedding based on the proposed label-sensitive and hard example mining strategies even in such cases that the face images undergo diverse changes of facial expressions. Moreover, our M-LSDML efficiently improves the performance thanks to more discriminative information exploited in M-LSDML from different face aging datasets, so that the learned feature similarity achieves robustness to diverse facial expressions.

I. Evaluation on ChaLearn

Lastly, we evaluated our LSDML on the apparent age estimation challenge dataset [49]. Since the ground-truth age labels of testing datasets are not available, we performed age estimation by utilizing the validation set for testing. Note that each face in the dataset was annotated by the mean apparent age and the standard deviation. We also introduced the Gaussian error [49] to conduct experiments on the apparent facial age estimation dataset for the evaluation protocol. The Gaussian error is computed by the following formula:

$$g = 1 - \exp\left(-\frac{(t-m)^2}{2\sigma^2}\right) \tag{11}$$

where t is the predicted age value, m is labeled mean age apparent age and σ is the ground-truth standard deviation.

In this experiment, we trained our LSDML by leveraging the training set from the ChaLearn dataset. We created KNN with the BIF feature as the baseline method. We also compared our LSDML with OHRANK +BIF and CS-LBFL. Accordingly, OHRANK utilizes hand-crafted feature BIF, and CS-LBFL automatically learns linear feature filters directly from image pixels. Table XI tabulates the MAEs and Gaussian errors [49], and Fig. 8 shows the CS curves of our approach compared with the state-of-the-arts, respectively. Compared with existing feature learning methods, we see that our LSDML achieves very competitive performance, which shows the discriminativeness of the learned feature similarity and robustness to large variations of varying facial expressions, diverse aspect ratios, partial occlusions and cluttered background. It is valuable to see that VGG (softmax, Exp) improves the performance by utilizing additional extremal face data

TABLE XI
COMPARISONS OF GAUSSIAN ERRORS OF OUR APPROACH WITH STATE-OF-THE-ARTS ON THE CHALEARN [49] VALIDATION SET

Method	Model Description	Gaussian Error	External Datasets
BIF [11]	BIF [11] + KNN	0.89	-
BIF [11]	BIF [11] + OHRANK [9]	0.55	-
VGG (softmax, Exp) [74]	Deep Expectation	0.51	-
VGG (softmax, Exp) [74]	Deep Expectation	**0.28**	D_6
VGG (softmax, Exp) [75]	with pretrained VGG-16 Face Net [64]	**0.28**	D_6
CS-LBFL [15]	Cost-Sensitive Local Binary Feature Learning	0.45	-
Best from DCNN [31]	deep convolutional neural networks	0.359	D_1, D_2, D_3
Cascaded-CNN [32]	with error correction	0.355	D_3, D_4,D_5
Cascaded-CNN [32]	with end-to-end finetuning	0.312	D_3, D_4,D_5
Cascaded-CNN [32]	with end-to-end finetuning and error correction	0.297	D_3, D_4,D_5
LSDML	with OHRANK [9]	0.37	-
M-LSDML	with OHRANK [9]	0.34	D_2,D_5
LSDML	with end-to-end finetuning [19]	0.328	-
M-LSDML	with end-to-end finetuning [19]	0.315	D_2,D_5

D_1-CASIA-WebFace [76], D_2-MORPH [46], D_3-AdienceFaces [46]
D_4-Images of Groups [77], D_5-FG-NET [22], D_6-IMDB-WIKL (https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/)

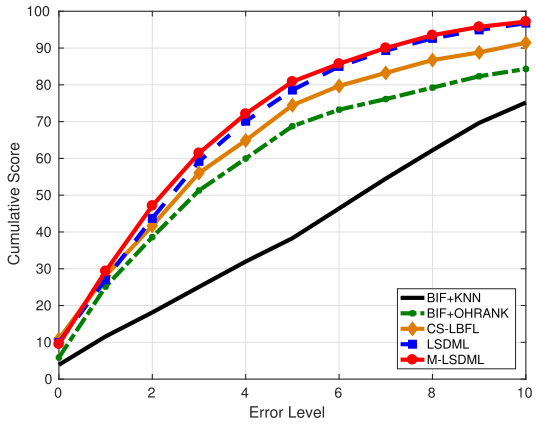


Fig. 8. The CS curves of our approach compared with different facial age estimation methods on the ChaLearn dataset.

and pre-trained model such as VGG-16 face net. Lastly, we compared our extension M-LSDML with LSDML and the results show that M-LSDML consistently outperforms LSDML, which demonstrates the effectiveness of utilizing multi-source cross-population datasets and the complementary correlation of different face aging datasets.

We also evaluated our proposed methods compared with DCNN [31] and Cascaded-CNN [32] on the ChaLearn [49] validation set. Since the source codes were not released, the results of the state-of-the-arts including [31] and [32] were directly cropped from the original papers. As the results are shown in Table XI, our methods outperform DCNN [31], which is because our model achieves the aging order information based on the proposed label-sensitive criterion. Moreover, compared with Cascaded-CNN [32] without end-to-end finetuning, our M-LSDML with OHRANK [9] obtains comparable performance by only utilizing D_2-MORPH [46] and D_5-FG-NET [22] that were captured in the controlled conditions. Moreover, we have evaluated our method by introducing an end-to-end technique. Specifically, we revised the proposed loss function by including the ordinal regression [19] at the top of the deep network, and then finetuned the network parameters. Having been integrated with an end-to-end finetuning technique [19], the performance based our M-LSDML has been improved by decreasing about 0.03 of Gaussian error, which shows the very competitive results with the best performance from Cascaded-CNN [32].

J. Discussion

The above experimental results suggest the following three observations:

1) The deep feature representations learned by our LSDML achieve better performance than those hand-crafted features. The reason is that our LSDML automatically learn age-adaptive and discriminative patterns directly from raw pixels, which performs strong robustness to diverse facial expressions, aspect ratios and cluttered background. Moreover, our model achieves the nonlinear relationship between face samples and exploits the label correlation for age classes in the learned feature similarity. Hence, better age estimation performance is obtained.
2) Our LSDML achieves better performance in comparisons to the deep metric learning methods which utilize random sampling method. The reason is two-fold: 1) our model aims to seek the feature similarity of face pairs with different age value gaps, which exploits the label correlation in the learned metric space, and 2) we learn to mine hard examples to preserve the local manifold structure of face samples in the newly learned subspace. In terms of the computational efficiency, our method exhibits fast convergence.
3) M-LSDML outperforms LSDML, which is because our M-LSDML aims to maximize the correlation for cross-population facial age datasets, which enhances the cross-population targets for training. As a result, the multi-source datasets reinforce to learn a more robust and discriminative feature similarity.

V. CONCLUSIONS AND FUTURE WORK

In this work, we have proposed a label-sensitive deep metric learning (LSDML) method for facial age estimation. Unlike existing hand-crafted features which require expert knowledge

by hand and the kernel-based metric learning methods which encounter the scalability problem, our LSDML leverages deep residual network to learn a series of nonlinear feature transformations, where the feature similarity is smoothly sensitive to the degree of age difference. To make the learned feature similarity more robust, we have proposed a hard example sampling method, which learns to select meaningful hard samples during the optimization process. Moreover, we have extended our proposed LSDML to M-LSDML by maximizing the correlation of cross-population multi-source datasets, so that the learned metric is more discriminative and robust. Experimental results on four benchmarking datasets show the effectiveness of our proposed approach. It is desirable to address facial age estimation with feed-back deep architectures for personalized face aging and jointly optimize the procedures of extracting age-related feature representation and predicting age in an end-to-end manner are are future works.

REFERENCES

[1] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.

[2] J. Lu, V. E. Liong, G. Wang, and P. Moulin, "Joint feature learning for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 7, pp. 1371–1383, Jul. 2015.

[3] G. Guo and C. Zhang, "A study on cross-population age estimation," in *Proc. CVPR*, 2014, pp. 4257–4263.

[4] X. Shu, J. Tang, H. Lai, L. Liu, and S. Yan, "Personalized age progression with aging dictionary," in *Proc. ICCV*, 2015, pp. 3970–3978.

[5] W. Wang *et al.*, "Recurrent face aging," in *Proc. CVPR*, 2016, pp. 2378–2386.

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[7] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[8] T. Wu, P. Turaga, and R. Chellappa, "Age estimation and face verification across aging using landmarks," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1780–1788, Dec. 2012.

[9] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. CVPR*, 2011, pp. 585–592.

[10] Y.-L. Chen and C.-T. Hsu, "Subspace learning for facial age estimation via pairwise age ranking," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 12, pp. 2164–2176, Dec. 2013.

[11] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. CVPR*, 2009, pp. 112–119.

[12] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.

[13] Z. Li, U. Park, and A. K. Jain, "A discriminative model for age invariant face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 2, pp. 1028–1037, Sep. 2011.

[14] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1955–1976, Nov. 2010.

[15] J. Lu, V. E. Liong, and J. Zhou, "Cost-sensitive local binary feature learning for facial age estimation," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5356–5368, Dec. 2015.

[16] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in *Proc. ACCV*, 2014, pp. 144–158.

[17] X. Liu *et al.*, "AgeNet: Deeply learned regressor and classifier for robust apparent age estimation," in *Proc. ICCVW*, 2015, pp. 258–266.

[18] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *Proc. WACV*, 2015, pp. 534–541.

[19] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. CVPR*, Jun. 2016, pp. 4920–4928.

[20] J. Weston, S. Bengio, and N. Usunier, "WSABIE: Scaling up to large vocabulary image annotation," in *Proc. IJCAI*, 2011, pp. 2764–2770.

[21] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. CVPR*, 2014, pp. 1386–1393.

[22] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 442–455, Apr. 2002.

[23] Y. Zhang and D.-Y. Yeung, "Multi-task warped Gaussian process for personalized age estimation," in *Proc. CVPR*, 2010, pp. 2622–2629.

[24] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *Proc. CVPR*, 2011, pp. 657–664.

[25] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578–584, Jun. 2008.

[26] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1178–1188, Jul. 2008.

[27] H. Liu, R. Ji, Y. Wu, and W. Liu, "Towards optimal binary code learning via ordinal embedding," in *Proc. AAAI*, 2016, pp. 1258–1265.

[28] Y. Dong, Y. Liu, and S. Lian, "Automatic age estimation based on deep learning algorithm," *Neurocomputing*, vol. 187, pp. 4–10, Apr. 2016.

[29] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. CVPRW*, 2015, pp. 34–42.

[30] Z. Hu, Y. Wen, J. Wang, M. Wang, R. Hong, and S. Yan, "Facial age estimation with age difference," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3087–3097, Jul. 2017.

[31] R. Ranjan *et al.*, "Unconstrained age estimation with deep convolutional neural networks," in *Proc. ICCV*, 2015, pp. 109–117.

[32] J.-C. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, and R. Chellappa, "A cascaded convolutional neural network for age estimation of unconstrained faces," in *Proc. BTAS*, 2016, pp. 1–8.

[33] H.-F. Yang, B.-Y. Lin, K.-Y. Chang, and C.-S. Chen, "Automatic age estimation from face images via deep ranking," in *Proc. BMVC*, 2015, pp. 55.1–55.11.

[34] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[35] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proc. NIPS*, 1993, pp. 737–744.

[36] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. CVPR*, 2005, pp. 539–546.

[37] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. CVPR*, 2014, pp. 1875–1882.

[38] J. Hu, J. Lu, Y.-P. Tan, and J. Zhou, "Deep transfer metric learning," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5576–5588, Dec. 2016.

[39] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. CVPR*, 2016, pp. 4004–4012.

[40] C. Huang, C. C. Loy, and X. Tang, "Local similarity-aware deep feature embedding," in *Proc. NIPS*, 2016, pp. 1262–1270.

[41] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. NIPS*, 2005, pp. 1473–1480.

[42] E. Arias-Castro. (Jan. 2015). "Some theory for ordinal embedding." [Online]. Available: https://arxiv.org/abs/1501.02861

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[44] I. Masi, A. T. Trân, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *Proc. ECCV*, 2016, pp. 579–596.

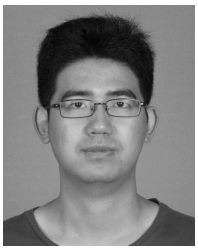
[45] H. Liu, J. Lu, J. Feng, and J. Zhou, "Group-aware deep feature learning for facial age estimation," *Pattern Recognit.*, vol. 66, pp. 82–94, Jun. 2017.

[46] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. FGR*, 2006, pp. 341–345.

[47] N. C. Ebner, M. Riediger, and U. Lindenberger, "FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation," *Behavior Res. Methods*, vol. 42, no. 1, pp. 351–362, 2010.

[48] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2170–2179, Dec. 2014.

[49] S. Escalera *et al.*, "Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *Proc. ICCVW*, 2015, pp. 243–251.
[50] D. E. King, "Dlib-ml: A machine learning toolki," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jul. 2009.
[51] Y. Jia *et al.* (Jun. 2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093
[52] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, vol. 9. 2010, pp. 249–256.
[53] G. Guo and G. Mu, "Human age estimation: What is the influence across race and gender?" in *Proc. CVPR*, 2010, pp. 71–78.
[54] J. Lu and Y.-P. Tan, "Cost-sensitive subspace learning for human age estimation," in *Proc. ICIP*, 2010, pp. 1593–1596.
[55] L. Miao, M. Liu, and D. Zhang, "Cost-sensitive feature selection with application in software defect prediction," in *Proc. ICPR*, 2012, pp. 967–970.
[56] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. CVPR*, 2013, pp. 2467–2474.
[57] R. Weng, J. Lu, G. Yang, and Y.-P. Tan, "Multi-feature ordinal ranking for facial age estimation," in *Proc. FG*, 2013, pp. 1–6.
[58] G. Guo and G. Mu, "A framework for joint estimation of age, gender and ethnicity on a large database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 761–770, 2014.
[59] K.-Y. Chang and C.-S. Chen, "A learning framework for age rank estimation based on face images with scattering transform," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 785–798, Mar. 2015.
[60] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. (Nov. 2016). "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size." [Online]. Available: https://arxiv.org/abs/1602.07360
[61] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[62] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, 2015, pp. 1–9.
[63] X. Wu, R. He, T. Tan, and Z. Sun. (Nov. 2015). "A light CNN for deep face representation with noisy labels." [Online]. Available: https://arxiv.org/abs/1511.02683
[64] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015, p. 6.
[65] X. Yang *et al.*, "Deep label distribution learning for apparent age estimation," in *Proc. ICCVW*, 2015, pp. 344–350.
[66] Z. Lei, M. Pietikäinen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 289–302, Feb. 2014.
[67] S. U. Hussain, T. Napoléon, and F. Jurie, "Face recognition using local quantized patterns," in *Proc. BMVC*, 2012, pp. 1–11.
[68] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "Ica with reconstruction cost for efficient overcomplete feature learning," in *Proc. NIPS*, 2011, pp. 1017–1025.
[69] S. Yan, H. Wang, X. Tang, and T. S. Huang, "Learning auto-structured regressor from uncertain nonnegative labels," in *Proc. ICCV*, 2007, pp. 1–8.
[70] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "A probabilistic fusion approach to human age prediction," in *Proc. CVPRW*, 2008, pp. 1–6.
[71] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "A ranking approach for human ages estimation based on face images," in *Proc. ICPR*, 2010, pp. 3396–3399.
[72] C. Li, Q. Liu, J. Liu, and H. Lu, "Learning distance metric regression for facial age estimation," in *Proc. ICPR*, 2012, pp. 2327–2330.
[73] G. Guo and X. Wang, "A study on human age estimation under facial expression changes," in *Proc. CVPR*, 2012, pp. 2547–2553.
[74] R. Rothe, R. Timofte, and L. Van Gool, "DEX: Deep expectation of apparent age from a single image," in *Proc. ICCVW*, 2015, pp. 252–257.
[75] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," in *Proc. IJCV*, 2016, pp. 1–14.
[76] D. Yi, Z. Lei, S. Liao, and S. Z. Li. (Nov. 2014). "Learning face representation from scratch." [Online]. Available: https://arxiv.org/abs/1411.7923
[77] A. C. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. CVPR*, 2009, pp. 256–263.



Hao Liu received the B.S. degree in software engineering from Sichuan University, China, in 2011, and the M.Eng. degree in computer technology from the University of Chinese Academy of Sciences, China, in 2014. He is currently pursuing the Ph.D. degree with the Department of Automation, Tsinghua University. His research interests include face alignment, facial age estimation, and deep learning.

Jiwen Lu received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and machine learning. He has authored or co-authored over 160 scientific papers in these areas, where 50 were the IEEE Transactions papers. He serves as an Associate Editor of *Pattern Recognition*, *Pattern Recognition Letters*, *Neurocomputing*, and IEEE ACCESS. He is an Elected Member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society and the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society. He was a recipient of the National 1000 Young Talents Plan Program in 2015.

Jianjiang Feng received the B.S. and Ph.D. degrees from the School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, China, in 2000 and 2007. From 2008 to 2009, he was a Post-Doctoral Researcher with the PRIP Laboratory, Michigan State University. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing. He is an Associate Editor of the *Image and Vision Computing*. His research interests include fingerprint recognition and computer vision.

Jie Zhou received the B.S. and M.S. degrees from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995. From 1995 to 1997, he was a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a Full Professor with the Department of Automation, Tsinghua University. In recent years, he has authored over 100 papers in peer-reviewed journals and conferences. Among them, over 40 papers have been published in top journals and conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and CVPR. His current research interests include computer vision, pattern recognition, and image processing. He received the National Outstanding Youth Foundation of China Award. He is an Associate Editor of the *International Journal of Robotics and Automation* and two other journals.